NEW SYSTEM FOR SME-WEBSITE-SCRAPER:

Our system consist of 3 containers:

1)Scraper

2)app-server

3)Database

Connections made between container , scraper connected to Database,app-server connected to Database

BLACK BOX VIEW:

1.When system get started using sloppy.json.The 3 containers will get started.

2.The scraper container access the urls from Database and schedule the crawling using spider.

3.The crawled data will be saved to Database

4.REST API will interact with the external DB instance and provide output in JSON format

5.Once the scraper container crawl the whole set of urls.It will compare the data crawled between previous run and present run ,calculate and show the difference in graph.This graph will be pushed to S3

WHITE BOX VIEW:

1.We will start the docker in sloppy.io with required environment variables.

2.When the server become online the appserver,database,scraper will become live.

3.In the scraper server their is a ENTRYPOINT set that will start the crawling.i.e when the scraper container become live the urls to crawl is taken from DB and the schedule to crawl (Crawling is scheduled at every interval of 2 min).

5.We generally use regular expression ,key text,Xpath search for getting data from site

6.Each time when spider crawl the data .The final data after filtering for duplicates,negatives it will get saved.

6.When a spider get closed the data will be saved to external DB instance.By using the connection made between Scraper and Database.

7.This Database will be given for processing by REST API framework

8.When a request come to the REST API framework .The framework will access the external db ,process the Request pass through the cases.Finally take the data from the db.

9.The result is given back using the REST API framework to the user.